Отчет о работе в семестре по предмету "Математический практикум"

Алина Терехова

9 декабря 2025 г.

В данной работе производился вывод асимптотики комбинированного метода UniXGraD-DoG + Smoothing technic для оптимизации негладких функций со стохастическим оракулом нулевого порядка с помощью алгоритмов оптимизации гладких функций с оракулами первого порядка.

1 Описание алгоритма UniXGrad-DoG

Схема алгоритма выглядит следующим образом:

Algorithm 1: U-DoG (UniXGrad-DoG) template

Input: Initial $x_0 \in \mathcal{K}$, iteration budget T, initial movement r_ϵ, step sizes $\{\eta_{x,t}, \eta_{y,t}\}$

Set $y_0 = x_0$

for $t = 0, 1, 2, \dots, T-1$ **do**

  Set $\alpha_t = \sum_{k=0}^{t} \bar{r}_k / \bar{r}_t$ and $\omega_t = \alpha_t \bar{r}_t$ for $\bar{r}_t = \max\limits_{k \le t} \max\{\|y_k - x_0\|, \|x_k - x_0\|, r_\epsilon\}$

  $x_{t+1} = \mathrm{Proj}_{\mathcal{K}}(y_t - \alpha_t \eta_{x,t} m_t)$ for $m_t \sim \mathcal{G}(\hat{z}_t)$ and $\hat{z}_t = \frac{\omega_t y_t + \sum_{k=0}^{t-1} \omega_k x_{k+1}}{\sum_{k=0}^{t} \omega_k}$

  $y_{t+1} = \mathrm{Proj}_{\mathcal{K}}(y_t - \alpha_t \eta_{y,t} g_t)$ for $g_t \sim \mathcal{G}(\hat{x}_t)$ and $\hat{x}_t = \frac{\omega_t x_{t+1} + \sum_{k=0}^{t-1} \omega_k x_{k+1}}{\sum_{k=0}^{t} \omega_k}$

end

return $\hat{x}_T$

Рис. 1: UniXGrad-DoG схема

Алгоритм оптимизации работает с выпуклыми, липшицевыми, гладкими функциями, определенными на замкнутом выпуклом множестве и имеющими несмещенный градиент.

Обозначения вводятся следующими:

- L - константа липшицевости
- β - константа гладкости
- $\mathcal{K}$ - замкнутое выпуклое множество, на котором определена f
- $\mathcal{G}$ - стохастический оракул, предоставляющий несмещенный градиент $\mathbb{E}\mathcal{G}(x) = \nabla f(x)$

Поясним принцип действия алгоритма с данных шаблоном:

- На вход схемы подается начальная точка x_0, сколько максимум итераций можно сделать T, грубая нижняя оценка на расстояние до решения r_ϵ, а также величины шагов для одновременного сдвига двух точек x и y: $\eta_{x,t}$ и $\eta y, t$, которые мы научимся подбирать позднее.
- В начале $y_0 = x_0$. На каждой итерации:
  - Получаем оценку на максимальное удаление от решения $\bar{r}_t = \max_{k \le t} \max\{||y_k - x_0||, ||x_k - x_0||, r_\epsilon\}$, т.е. это нижняя оценка на диаметр нашего множества.
  - Высчитываем коэффициент $\alpha_t = \sum_{k=0}^{t} \bar{r}_k / \bar{r}_t$ и вес $w_t = \sum_{k=0}^{t} \bar{r}_k$. Вес в дальнейшем мы будем использовать, чтобы получать взвешенное среднее всех точек, куда мы когда-либо попадали: так как он возрастает все время, то новые точки вносят больший вклад в подсчет ответа. Коэффициент α_t растет примерно линейно с t по прошествии большого числа итераций. Он используется для того, чтобы уверенней делать шаги со временем.
  - Мы определяем средневзвешенное значение $\hat{z}_t = \frac{\omega_t y_t + \sum_{k=0}^{t-1} \omega_k x_{k+1}}{\sum_{k=0}^{t} \omega_k}$ точек y_k, и с помощью оценки градиента в этой точке делаем шаг из точки y_t, записывая результат в x_{t+1}. Проекция нужна для того, чтобы случайно не выпрыгнуть за пределы исследуемого множества.
  - То же самое делаем для обновления y_{t+1}. Можно сказать, что y_{t+1} - это оценка на x_{t+2}.

Осталось сказать пару слов про обновление шагов и про ограничения на шум градиента.

Мы работаем далее только со случаем ограниченного множества, так что шаги определяются следующим образом:

$$\eta_{x,t} = \eta_{y,t} = \frac{\bar{r}_t}{\sqrt{1 + Q_{t-1}}}$$

где $Q_{t-1} = \sum_{k=0}^{t-1} \alpha_k^2 ||g_t - m_t||^2$ - оценка того, насколько шумно сейчас выдается градиент. Чем выше эта величина, тем меньше делается шаг.

Что касается шума, мы анализируем ситуацию, когда он ограничен сверху некоторой константой σ, т.е. $||\mathcal{G}(x) - \nabla f(x)|| \le \sigma$. Со всеми введенными обозначениями в аппендиксе статьи выводится асимптотика:

$$\tilde{O}(\frac{\beta d_0^2}{T^2} + \frac{\sigma d_0}{\sqrt{T}})$$

для скорости сходимости алгоритма (так как мы имеем дело со случаем ограниченного множества, тут d_0 означает нижнюю оценку на его диаметр)

2 Описание Smoothing scheme

Пусть у нас имеется задача оптимизации функции f на выпуклом множестве Q. Сама фунция f выпукла на множестве $Q_\gamma = Q + B_2^d(\gamma)$, где последнее слагаемое - это евклидов шар радиуса γ с центром в нуле. Далее, мы делаем следующие предположения:

- Функция f - липшицева на Q_γ. Обозначим константу, для совместимости с предыдущей статьей, как L. (при этом она липшицева по p-норме; если p=2 обозначим константу за L_2)
- У нас есть оракул нулевого порядка, выдающий значение функции с некоторым ограниченным шумом Δ.

Из этих предположений выводятся следующие утверждения для сглаженной функции $f_\gamma = \mathbb{E}f(x+u)$, где u - случайный вектор, равномерно распределенный в шаре $B_2^d(\gamma)$. Оговоримся, что градиент данной функции аппроксимируется, как $\nabla f_\gamma(x,e) = d\frac{f(x+\gamma e)-f(x-\gamma e)}{2\gamma}e$, где e - случайный вектор, равномерно распределенный по сфере единичного радиуса с центром в нуле.

- f_γ - L-липшицева на Q_γ по норме p
- f_γ - гладкая с константой $\beta = \frac{\sqrt{d}L}{\gamma}$, где $\frac{1}{q}+\frac{1}{p}=1$

$$||\nabla f_\gamma(y) - \nabla f_\gamma(x)||_q \leq \beta||x-y||_p \tag{1}$$

- Для аппроксимации градиента f_γ справедливо:

$$\sigma^2 = \mathbb{E}_e[||\nabla f(x,e)||_q^2] \leq \kappa(p,d)(dM_2^2 + \frac{d^2\Delta^2}{\gamma^2}) \tag{2}$$

где $\kappa(p,d) = O(\min\{q, \ln d\}d^{\frac{d}{q-1}})$, d - размерность пространства, откуда действует f.

Теперь опишем суть метода.

Пусть у нас есть алгоритм $\mathbf{A}(\beta,\sigma^2)$, который решает задачу оптимизации функции f на Q_γ, которая одновременно:

- Гладкая с константой β и для нее выполнено $\forall x,y \in Q_\gamma$

$$||\nabla f(x) - \nabla f(y)||_q \leq \beta||x-y||_p$$

- Имеет стохастический оракул 1-го порядка, который зависит от случайной величины η и выдает несмещенный градиент с ограниченной дисперсией:

$$\mathbb{E}_\eta[||\nabla_x f(x,\eta) - \nabla f(x)||] \leq \sigma^2$$

т.е. выполнено (2)

Идея в том, чтобы использовать этот алгоритм для оптимизации функции f_γ с $\gamma = \frac{\varepsilon}{2L_2}$, $\eta = e$, $\nabla_x f(x,\eta) = \nabla_\gamma f(x,e)$. Таким образом, мы получим алгоритм оптимизации негладкой стохастической функции с оракулом нулевого порядка.

Если алгоритм $\mathbf{A}$ требовал $N(\beta,\varepsilon)$ успешных итераций и $T(\beta,\sigma^2,\varepsilon)$ вызовов стохастического оракула 1-го порядка, то для решения новой задачи потребуется:

$$N(\frac{2\sqrt{d}LL_2}{\varepsilon}, \varepsilon)$$

$$2T(\frac{2\sqrt{d}LL_2}{\varepsilon}, 2\kappa(p,d)dM_2^2, \varepsilon)$$

3 Изменения в доказательстве асимптотики U-DoG

Поскольку далее в работе нам потребуется оценка на матожидание $f(\hat{x}_t) - f(x^*)$, получим его, используя доказательство асимптотики U-DoG.

Возьмем матожидание от следующего выражения:

$$f(\hat{x}_t) - f(x^*) \le 360 s^{3/2} \frac{\beta(\overline{r}_{t+1} + \overline{d}_{t+1})^2}{(\sum_{k=0}^t \overline{r}_k / \overline{r}_{t+1})^2}$$

$$+10 \frac{(\overline{r}_{t+1} + \overline{d}_{t+1})(\sqrt{max\{G_{y,t}, Q_t\}} - s\sqrt{Q_t})}{(\sum_{k=0}^t \overline{r}_k / \overline{r}_{t+1})^2}$$

$$+20 \frac{(\overline{r}_{t+1} + \overline{d}_{t+1})\sqrt{(t+1)^3 V_t + (\theta_{t+1,\delta}\sigma)^2}}{(\sum_{k=0}^t \overline{r}_k / \overline{r}_{t+1})^2}$$

Здесь три слагаемых, следовательно, в силу линейности матожидания:

$$\mathbb{E}[f(\hat{x}_t) - f(x^*)] \le \mathbb{E}[A] + \mathbb{E}[B] + \mathbb{E}[C]$$

Разберем по порядку:

1) A

$$\mathbb{E}[A] = 360 s^{3/2} \beta \cdot \mathbb{E}[\frac{(\overline{r}_{t+1} + \overline{d}_{t+1})^2}{(\sum_{k=0}^t \overline{r}_k / \overline{r}_{t+1})^2}]$$

Заметим, что $\overline{r}_{t+1} + \overline{d}_{t+1} \le 6d_0$

$$\mathbb{E}[A] \le 360 s^{3/2} \beta \cdot 36 d_0^2 \cdot \mathbb{E}[\frac{1}{(\sum_{k=0}^t \overline{r}_k / \overline{r}_{t+1})^2}]$$

Знаменатель ведет себя, как $\Omega(\frac{t^2}{log^2 t})$

$$\mathbb{E}[A] \le 360 s^{3/2} \beta \cdot 36 d_0^2 \cdot O(\frac{log^2 t}{t^2}) = O(\frac{s^{3/2}\beta d_0^2 log^2 t}{t^2})$$

2) B

$$\mathbb{E}[B] = 10\mathbb{E}[\frac{(\overline{r}_{t+1} + \overline{d}_{t+1})(\sqrt{max\{G_{y,t}, Q_t\}} - s\sqrt{Q_t})}{(\sum_{k=0}^t \overline{r}_k / \overline{r}_{t+1})^2}] \le 60 d_0 \mathbb{E}[\frac{\sqrt{max\{G_{y,t}, Q_t\}} - s\sqrt{Q_t}}{(\sum_{k=0}^t \overline{r}_k / \overline{r}_{t+1})^2}]$$

Заметим, что знаменатель вновь есть $\Omega(\frac{t^2}{log^2 t})$ Проанализируем числитель. Заметим, что в случае ограниченного множества, на котором исследуется функция: $G_{y,t} = 1 + Q_t$, таким образом:

$$\sqrt{max\{G_{y,t}, Q_t\}} - s\sqrt{Q_t} = \sqrt{1+Q_t} - s\sqrt{Q_t} \le (1-s)\sqrt{1+Q_t} \le (1-s)\sqrt{1+\sigma^2 t \alpha_t^2} = O((1-s)\sigma\sqrt{t}\alpha_t)$$

Подставим всё в предыдущее соотношение:

$$\mathbb{E}[B] \le 60 d_0 \cdot O(\frac{(1-s)\sigma \log t}{\sqrt{t}}) = O(\frac{(1-s)\sigma d_0 \log t}{\sqrt{t}})$$

3) C

$$\mathbb{E}[C] = 20\mathbb{E}[\frac{(\overline{r}_{t+1} + \overline{d}_{t+1})\sqrt{(t+1)^3 V_t + (\theta_{t+1,\delta}\sigma)^2}}{(\sum_{k=0}^t \overline{r}_k / \overline{r}_{t+1})^2}] \le 120 d_0 \mathbb{E}[\frac{\sqrt{(t+1)^3 V_t + (\theta_{t+1,\delta}\sigma)^2}}{(\sum_{k=0}^t \overline{r}_k / \overline{r}_{t+1})^2}]$$

Вновь проанализируем числитель. Заметим, что так как шум мы считаем ограниченным, то $V_t \le 2\sigma^2$.

$$\sqrt{O(t^3\sigma^2 + \sigma \log^2 \log t)} = O(t^{\frac{3}{2}}\sigma + \sqrt{\sigma}\log\log t)$$

Тогда для всего выражения:

$$\mathbb{E}[C] \le O(\frac{d_0 \sigma \log^2 t}{\sqrt{t}} + \frac{d_0\sqrt{\sigma}\log^2 t \log\log t}{t^2})$$

Собираем все вместе и получаем:

$$\mathbb{E}[f(\hat{x}_t) - f(x^*)] \le O(\frac{s^{3/2}\beta d_0^2 log^2 t}{t^2} + \frac{(1-s)\sigma d_0 \log t}{\sqrt{t}} + \frac{d_0\sigma \log^2 t}{\sqrt{t}} + \frac{d_0\sqrt{\sigma}\log^2 t \log\log t}{t^2})$$

4 Вывод асимптотики комбинированного метода

Мы имеем:

$$\mathbb{E}[f(\hat{x}_t) - f(x^*)] \leq O(\frac{s^{3/2}\beta d_0^2 log^2 t}{t^2} + \frac{(1-s)\sigma_B d_0 \log t}{\sqrt{t}} + \frac{d_0 \sigma_B \log^2 t}{\sqrt{t}} + \frac{d_0\sqrt{\sigma_B}\log^2 t \log\log t}{t^2})$$

Это дает для некоторых констант:

$$\mathbb{E}[f(\hat{x}_t) - f(x^*)] \leq C_1(\frac{(1-s)\sigma_B d_0 \log t}{\sqrt{t}} + \frac{d_0 \sigma_B \log^2 t}{\sqrt{t}}) + C_2(\frac{s^{3/2}\beta d_0^2 log^2 t}{t^2} + \frac{d_0\sqrt{\sigma_B}\log^2 t \log\log t}{t^2})$$

Для того, чтобы все решение было менее $\frac{\varepsilon}{2}$ достаточно требовать, чтобы каждое из слагаемых не превосходило $\frac{\varepsilon}{4}$. Запишем это условие и оценим значения N и B для такой точности:

$$C_1(\frac{(1-s)\sigma_B d_0 \log N}{\sqrt{N}} + \frac{d_0\sigma_B \log^2 N}{\sqrt{N}}) \leq \frac{\varepsilon}{4} \;\Rightarrow\; \sqrt{N} \geq \frac{4C_1 d_0 \sigma_B \log N(1 - s + \log N)}{\varepsilon}$$

Подставим σ_B:

$$\sqrt{N} \geq \frac{4C_1 d_0 \sigma\sqrt{\lambda(p,d)}\log N(1-s+\log N)}{\sqrt{B}\varepsilon}$$

Заметим, что $\exists N_0 : \forall N \geq N_o$ выполнено $\log N \geq \log\frac{16C_1^2 d_0^2\sigma^2}{\varepsilon^2 B}$

$$\Rightarrow N \geq \frac{16C_1^2 d_0^2 \sigma^2 \lambda(p,d)\log\frac{16C_1^2 d_0^2\sigma^2}{\varepsilon^2 B}(1-s+\log\frac{16C_1^2 d_0^2\sigma^2}{\varepsilon^2 B})^2}{\varepsilon^2 B}$$

С другой стороны, из второго слагаемого получаем:

$$C_2(\frac{s^{3/2}\beta d_0^2 log^2 t}{t^2} + \frac{d_0\sqrt{\sigma_B}\log^2 t\log\log t}{t^2}) \leq \frac{\varepsilon}{4} \;\Rightarrow\; N^2 \geq 4C_2\frac{d_0\log^2 N}{\varepsilon}(s^{3/2}\beta d_0 + \sqrt{\sigma_B}\log\log N)$$

То же самое рассуждение применимо тут: $\exists N_0 : \forall N \geq N_0 \; \log N \geq \log\frac{d_0}{\varepsilon}$ и подставим σ_B:

$$N^2 \geq 4C_2\frac{d_0\log^2\frac{d_0}{\varepsilon}}{\varepsilon}(s^{3/2}\beta d_0 + \sqrt{\sigma}\frac{\lambda(p,q)^{1/4}}{B^{1/4}}\log\log\frac{d_0}{\varepsilon})$$

$$N \geq 2C_2^{1/2}\frac{d_0^{1/2}\log\frac{d_0}{\varepsilon}}{\varepsilon^{1/2}}(s^{3/2}\beta d_0 + \sqrt{\sigma}\frac{\lambda(p,q)^{1/4}}{B^{1/4}}\log\log\frac{d_0}{\varepsilon})^{1/2}$$

Значит, итоговая оценка на N будет максимумом из этих двух величин:

$$N \geq \max\{2C_2^{1/2}\frac{d_0^{1/2}\log\frac{d_0}{\varepsilon}}{\varepsilon^{1/2}}(s^{3/2}\beta d_0 + \sqrt{\sigma}\frac{\lambda(p,q)^{1/4}}{B^{1/4}}\log\log\frac{d_0}{\varepsilon})^{1/2}, \frac{16C_1^2 d_0^2\sigma^2\lambda(p,d)\log\frac{16C_1^2 d_0^2\sigma^2}{\varepsilon^2 B}(1-s+\log\frac{16C_1^2 d_0^2\sigma^2}{\varepsilon^2 B})^2}{\varepsilon^2 B}\}$$

Получим оценку на B. Для начала рассмотрим второй элемент максимума:

$$N \geq \frac{K_1\log\frac{A}{B}(K_2 + \log\frac{A}{B})^2}{B} \geq \frac{K_1(\log\frac{A}{B})^3}{B}$$

Второй элемент:

$$N \geq P_1(P_2 + \frac{P_3}{B^{1/4}})^{1/2} \geq \frac{P_1 P_3^{1/2}}{B^{1/8}}$$

Балансировка:

$$\frac{P_1 P_3^{1/2}}{B^{1/8}} = \frac{K_1(\log\frac{A}{B})^3}{B} \;\Rightarrow\; B^{7/8} = \frac{K_1}{P_1 P_3^{1/2}}(\log\frac{A}{B})^3 \leq \frac{K_1}{P_1 P_3^{1/2}}(\log A)^3$$

$$B^{7/8} \leq \frac{8C_1^2 d_0^{3/2}\sigma^{7/4}\lambda^{7/8}(p,d)}{C_2^{1/2}\log\frac{d_0}{\varepsilon}(\log\log\frac{d_0}{\varepsilon})^{1/2}\varepsilon^{3/2}} \Rightarrow B = \tilde{O}(\frac{d_0^{12/7}\sigma^2}{\varepsilon^{12/7}})$$

Значит, для числа итераций:

$$N \geq \max\{\tilde{O}(\frac{d_0^{1/2}}{\varepsilon^{1/2}}(s^{3/2}\beta d_0 + \frac{\sigma^{1/2}}{B^{1/4}})^{1/2}), \tilde{O}(\frac{d_0^2\sigma^2}{\varepsilon^2 B})\})$$

$$N \geq \max\{\tilde{O}(T_1), \tilde{O}(T_2)\}$$

$$T_1 = \frac{d_0^{1/2}}{\varepsilon^{1/2}}(s^{3/2}\beta d_0 + \frac{\varepsilon^{3/7}}{d_0^{3/7}})^{1/2} = \sqrt{\frac{s^{3/2}\beta d_0^2}{\varepsilon} + (\frac{d_0}{\varepsilon})^{4/7}}$$

$$T_2 = (\frac{d_0}{\varepsilon})^{2/7}$$

Заметим, что $T_2 \leq T_1$ при любых значениях входящих переменных (т.к. все они положительны)

$$N = \tilde{\Omega}(\sqrt{\frac{s^{3/2}\beta d_0^2}{\varepsilon} + (\frac{d_0}{\varepsilon})^{4/7}}) = \tilde{\Omega}(\frac{s^{3/4}\beta^{1/2}d_0}{\sqrt{\varepsilon}})$$

Подстваим оценку $\beta = \frac{\sqrt{d}L}{\gamma} = \frac{2d^{1/2}LL_2}{\varepsilon}$

$$N = \tilde{\Omega}(\frac{s^{3/4}d^{1/4}\sqrt{LL_2}}{\varepsilon}d_0)$$

Тогда оценка для числа оракульных вызовов:

$$T = N \cdot B = \tilde{\Omega}(\frac{s^{3/4}d^{1/4}L^{1/2}L_2^{1/2}d_0^{19/7}\sigma^2}{\varepsilon^{19/7}})$$

5 Анализ осимптотики

В ходе работы была проанализирована асимптотика комбинированного метода UniXGraD-DoG и Smoothing schene. Полученная оценка на число итераций выглядит правдоподобно по сравнению с данными, характерными для задач рассматриваемого класса, однако оценка на размер батча требует более тщательной проверки. Дальнейший этап работы - это попытаться протестировать реальные функции и посмотреть на реальные зависимости для размера батча. Сравнение их с теоретически полученными позволит проверить корректность расчетов.